Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **21,** 3787–3793.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstraele, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, P., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34,** 267–273.

Reimers, M., and Carey, V. J. (2006). Bioconductor: An open source framework for bioinformatics and computational biology. *Methods Enzymol.* **411,** 119–134.

Quackenbush, J. (2003). Genomics. Microarrays: Guilt by association. *Science* **302,** 240–241.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A., and Quackenbush, J. (2006). TM4 microarray software suite. *Methods Enzymol.* **411,** 134–193.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470.

Shaffer, J. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.* **46,** 561–584.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *SAGMB.* **3,** Article 3.

Whetzel, P. L., Parkinson, H., and Stoecket, C. (2006). Using ontologies to annotate microarray experiments. *Methods Enzymol.* **411,** 325–339.

Young, A., Whitehouse, N., Cho, J., and Shaw, C. (2005). OntologyTraverser: An R package for GO analysis. *Bioinformatics* **21,** 275–276.

Yue, L., and Reisdorf, W. C. (2005). Pathway and ontology analysis: Emerging approaches connecting transcriptome data and clinical endpoints. *Curr. Mol. Med.* **5,** 11–20.

Further Reading

Eisen, M. B., Spellman, P. T., Browm, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.

Masys, D., Welsh, J., Fink, J. L., Gribskov, M., Klacansky, I., and Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* **17,** 319–326.

[19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis

By TANYA BARRETT and RON EDGAR

Abstract

The Gene Expression Omnibus (GEO) repository at the National Center for Biotechnology Information archives and freely distributes high-throughput molecular abundance data, predominantly gene expression data generated by DNA microarray technology. The database has a

0076-6879/06 $35.00
DOI: 10.1016/S0076-6879(06)11019-8

flexible design that can handle diverse styles of both unprocessed and processed data in a Minimum Information About a Microarray Experiment-supportive infrastructure that promotes fully annotated submissions. GEO currently stores about a billion individual gene expression measurements, derived from over 100 organisms, submitted by over 1500 laboratories, addressing a wide range of biological phenomena. To maximize the utility of these data, several user-friendly web-based interfaces and applications have been implemented that enable effective exploration, query, and visualization of these data at the level of individual genes or entire studies. This chapter describes how data are stored, submission procedures, and mechanisms for data retrieval and query. GEO is publicly accessible at http://www.ncbi.nlm.nih.gov/projects/geo/.

Purpose and Scope of the Gene Expression Omnibus (GEO)

The postgenomic era has led to a multitude of high-throughput methodologies that generate massive volumes of gene expression data. The GEO repository was established by National Center for Biotechnology Information (NCBI) in 2000 to house and distribute these data to the public with no restrictions or login requirements (for more information, please read the GEO data disclaimer[1]). The primary role of GEO is data archiving, functioning as a hub for data deposit, and retrieval (Barrett *et al.*, 2005; Edgar *et al.*, 2002). ArrayExpress (Brazma *et al.*, 2006) serves a similar function.

GEO is currently the largest, fully public gene expression resource. At the time of writing, the database holds over 80,000 samples, comprising approximately a billion individual expression measurements, 13 million gene expression profiles, for over 100 organisms, submitted by almost 1500 laboratories. These data address a very broad diversity of biological themes, including disease, development, evolution, metabolics, toxicology, immunity, ecology, and transgenesis. Most data are provided by the research community in compliance with grant or journal provisos that require microarray data to be made available in a public repository, with the objective being to facilitate independent evaluation of results, reanalysis, and full access to all parts of the study (Ball *et al.*, 2004).

Data types currently stored include, but are not limited to, cDNA and oligonucleotide microarrays that examine gene expression, serial analysis of gene expression (SAGE), massively parallel signature sequencing, array comparative genomic hybridization, chromatin–immunoprecipitation on arrays studies, and peptide profiling techniques such as tandem mass

[1] http://www.ncbi.nlm.nih.gov/projects/geo/info/disclaimer.html.

spectrometry (MS/MS). In keeping with the theme of the book, this chapter focuses on gene expression data generated by DNA microarrays.

Although primarily a data storage and retrieval facility, it was clear early on that the resource must also enable effective searching and data mining as means to identify entries of interest. Consequently, several user-friendly web-based query tools have been developed to assist even those unfamiliar with microarray technology to effectively explore and analyze GEO data. However, it is important to realize that GEO is not intended to be used as a laboratory information management system or a pre-/first-analysis environment, as data submitted to GEO are generally processed data that form the basis for discussion in accompanying manuscripts.

This chapter explains the database design for storage of microarray information, how to submit data, and how to effectively retrieve and examine information in the GEO database.

Structure

The GEO database architecture is designed for the efficient capture, storage, and retrieval of heterogeneous sets of high-throughput molecular abundance data. The structure is sufficiently flexible to accommodate evolving state-of-the-art technologies. There are many different varieties of microarray technology, and researchers use a wide assortment of hardware and software packages to generate and process data. Consequently, data have many different styles and comprise varying content. For example, the sequences on an array may be described by multiple attributes, including gene symbols, GenBank accession numbers, clone identifiers, ontology categories, and feature coordinates, to name a few. Similarly, hybridization data may contain many types of supporting measurements and calculations that supplement final expression values. Importantly, expression data are worthless unless complemented with comprehensive contextual biological details and data analysis methodologies under which they were generated. GEO was built with all these considerations in mind and has an open, adaptable design that can handle variety and a Minimum Information About a Microarray Experiment (MIAME)-supportive (Brazma *et al.*, 2001) infrastructure that promotes fully annotated submissions. Extensive technical details regarding database design and data flow are beyond the scope of this chapter, but it helps to understand that data and metadata are stored separately within the database. The versatility of GEO is largely attributed to the fact that tabular data are not fully granulated in the core database but instead are treated as ''blobs,'' that is, compressed text tab-delimited tables that may contain any number of rows or columns. Data in selected columns are extracted to a secondary database and used in

subsequent indexing and query applications. Descriptive or informative metadata are fully normalized in the schema as needed.

Submitter-Supplied Data

Data supplied by submitters are stored as three main entities in a MSSQL server relational database.

Platform: Includes a summary description of the array and a data table defining the array template. Each row in the table corresponds to a single element and includes sequence annotation and tracking information as provided by the submitter.

Sample: Includes a description of the biological source and the experimental protocols to which it was subjected and a data table containing hybridization measurements for each element on the corresponding platform.

Series: Defines a set of related samples considered to be part of a study and describes the overall study aim and design.

Each of these three objects is assigned an accession number that may be used to cite and retrieve the records. In addition to sample data tables and descriptive information, accompanying supplementary files such as original microarray scan images or preprocessed quantification data are accepted and stored on an FTP site with database links.

GEO-Constructed Data Sets

Despite the variability in the style and content of incoming data, a common set of salient information is submitted:

- sequence identity tracking information for each feature on the array
- normalized hybridization measurements
- a description of the biological source used in each hybridization.

Using a combination of automated data extraction and manual curation, this information is rendered into an upper-level unit called a GEO DataSet (Fig. 1). A DataSet represents a collection of similarly processed, experimentally related sample hybridizations and provides a coherent synopsis for a study. Samples within a DataSet are further categorized according to experimental variables, for example, they are organized by gender and disease state.

A DataSet provides two separate perspectives of data.

1. An *experiment-centered* rendering that encapsulates the whole study. This information is presented as a "DataSet record." DataSet records comprise a synopsis of the experiment, a breakdown of the experimental

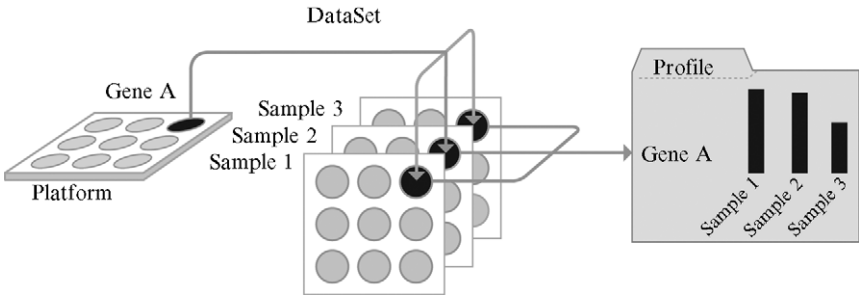

FIG. 1. Schematic diagram of relationships among GEO platform, sample, DataSet, and profiles. For each gene on a platform, multiple sample measurement values are generated. Related samples constitute a DataSet from which multiple gene expression profile entities are generated.

variables, access to auxiliary objects, several data display and analysis tools, and download options (Fig. 2).

2. A *gene-centered* rendering that presents quantitative gene expression measurements for one gene across a DataSet. This information is presented as a "GEO Profile." A GEO Profile comprises gene identity annotation, the DataSet title, links to auxiliary information, and a chart depicting the expression level of that gene across each sample in the DataSet (Fig. 3). The following section describes more information on interpreting GEO profile charts.

DataSets enable transformation of diverse styles of submitted data such that they are readily accessible in a uniform format upon which to base downstream data analysis tools.

Interpreting GEO Profiles Charts

GEO profile charts track the expression behavior of one gene across all samples in a DataSet. Several categories of information are presented in GEO profile charts: expression measurement values, expression measurement rankings, and an outline of the experimental design and variables (Fig. 3).

The *value* data (red bars, scale at the left side of the chart shown in Fig. 3) are extracted from the "VALUE" column of corresponding sample records from which the DataSet is composed. All sample data tables include this column, which contains the final normalized expression level measurements as supplied by the submitter. Other than to log transform single-channel expression counts for graphic visualization, no additional processing is applied by GEO to *value* data.

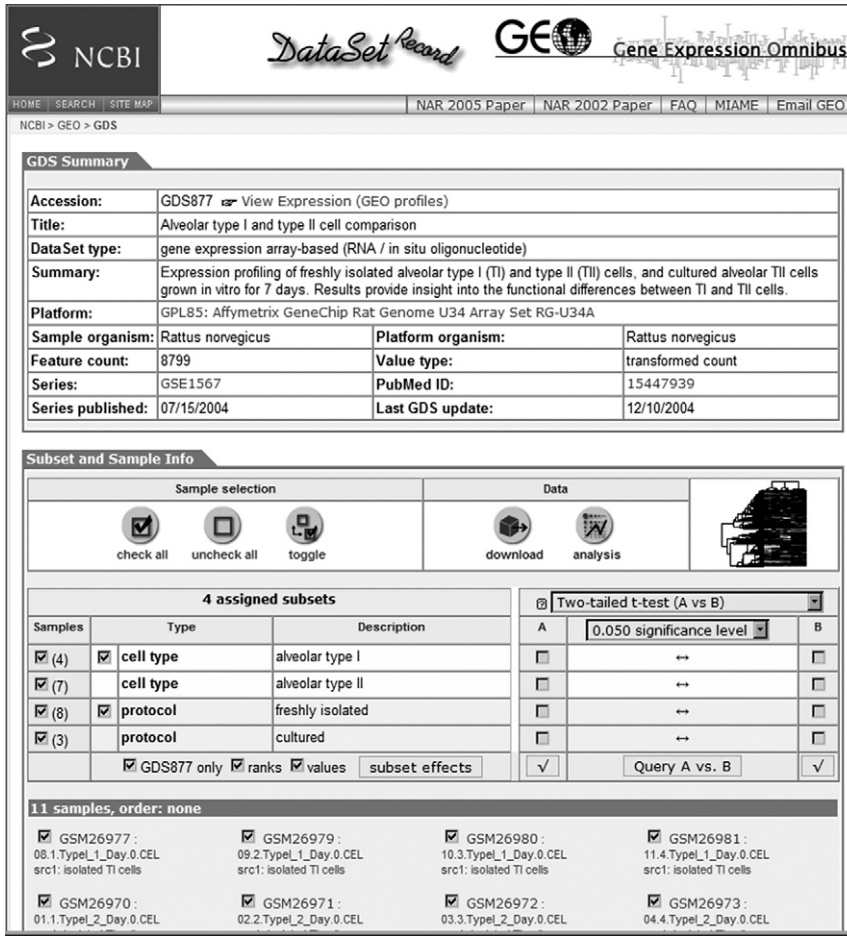

FIG. 2. Screen shot of a typical DataSet record GDS877 (Gonzalez *et al*., 2005). The record includes a summary of the experiment, links to related records and publications, subset designations and classifications, download options, and access to mining features such as cluster heat maps and ''Query group A vs B'' tool.

An important point to consider is that there is no standard unit for gene expression; because a very wide variety of technologies, software packages, and algorithms generate these data, the *values* should be considered arbitrary units. Consequently, it is inadvisable to attempt to draw direct comparisons between expression values in unrelated DataSets. However, it can be assumed that the *value* measurements of each sample within a DataSet

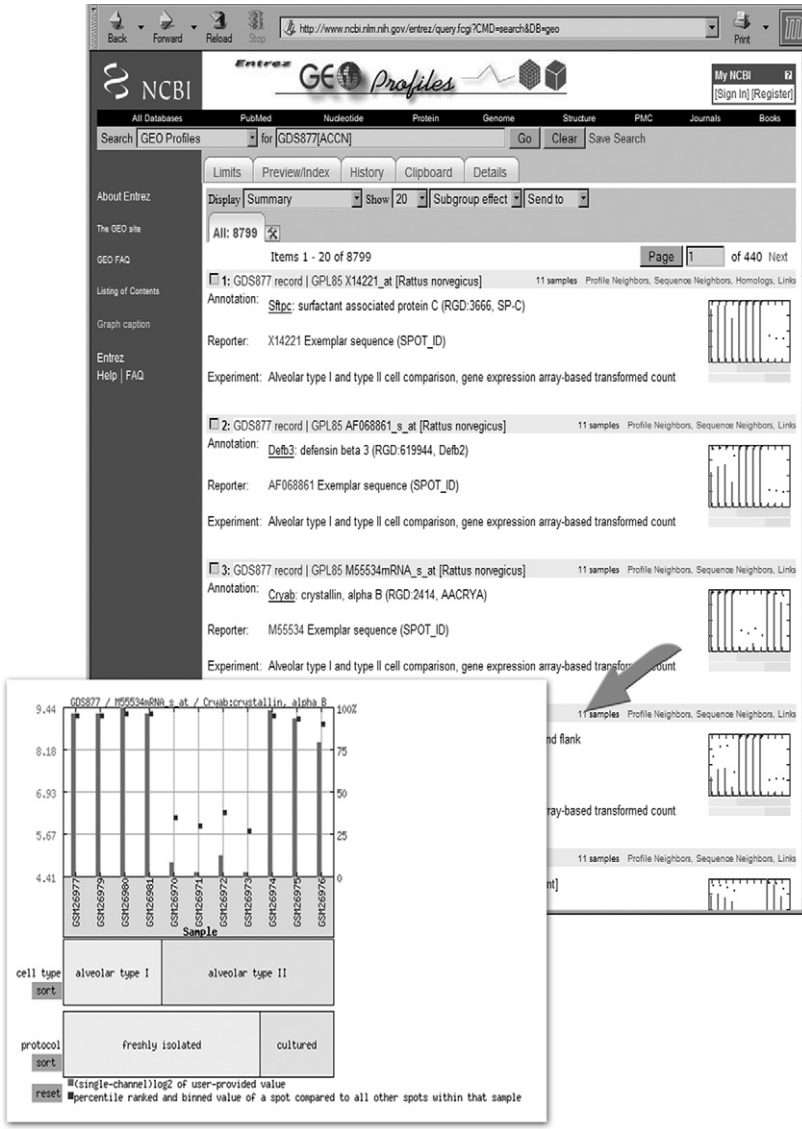

FIG. 3. Screen shot of Entrez GEO profile retrieval results; each entity includes sequence identifier and DataSet information and a thumbnail profile image. Links to other Entrez databases or related profiles are provided above the thumbnail image. The expanded profile chart depicts values (bars) and rank (squares) information for the crystallin gene across each sample in GEO DataSet GDS877 (Gonzalez *et al.*, 2005). Experimental subset groupings are reflected in labels at the foot of the chart.

are comparable and have been calculated in an equivalent manner, that is, considerations such as background processing and normalization/scaling are consistent across the DataSet. The ''Value distribution'' box and whisker plots available on DataSet records allow users to easily evaluate how well distributed, and thus comparable, the sample *values* within a DataSet are.

In addition to the *value* profile display for individual genes, most DataSets also provide a *rank* percentile view (blue squares, scale on the right side of the chart shown in Fig. 3). Ranks provide an indication of the expression level of that gene compared to all other genes on that array. Ranks are calculated as follows: (i) the total number of genes in the sample is divided to 100 bins such that there are n genes per bin; (ii) genes are sorted by *value*, and (iii) the lowest n genes are assigned to the first bin, subsequent n genes to the next bin, and so on. Binning is rather sensitive to local (sample) distribution and global (DataSet) normalization. It is therefore useful to note if a gene displays the same pattern of behavior in both value and rank space, as a disparity in trends can indicate that data are not normalized or the existence of other effects, such as nonspecific hybridization.

Currently, faded data points are specific to Affymetrix technology (this mode of display will likely be applied to other technology types in the future). They indicate where the Affymetrix algorithms have assigned a ''Detection call = absent'' to an expression signal. An absent call can be assigned for two reasons: either the detected signal was so low that the transcript was deemed not to be present or stray cross-hybridization was detected, in which case the signal is deemed unreliable for that transcript.

Bars at the horizontal foot of the chart provide experiment annotation and contextual information about the gene expression profile under review. The ''sort'' button allows users to resort the samples in the DataSet according to a particular experimental parameter, thus assisting visualization of expression trends in experiments with complex design.

Submission

The GEO database is a MIAME-supportive infrastructure; the MIAME guidelines outline the minimal information that should be provided to allow unambiguous interpretation of microarray experiment data (Brazma *et al.*, 2001). While the submission procedures promote MIAME compliance, ultimately it is the submitters' responsibility to ensure that their data are sufficiently well annotated. Large volumes of contextual information may be provided, including the cell or tissue type, characteristics of the organism

(e.g., species, age, sex, disease state) from which the sample was isolated, comprehensive explanations of the perturbations that the cells or organisms were subjected to, sample isolation and preparation protocols, data processing and normalization strategies, and more.

There are several ways in which data may be deposited with GEO. Deciding which method to use depends on the amount of data to be submitted, what format data are in already, and the level of computational expertise of the submitter. Regardless of the submission method, the final GEO records look the same and contain equivalent information.

Web deposit: The web submission process is designed for the quick and easy deposit of individual records by occasional submitters. This route comprises a set of interactive web forms that provide a simple step-by-step procedure for deposit of data tables and accompanying descriptive information.

Batch direct deposit using Simple Omnibus Format in Text (SOFT) format: SOFT is a simple line-based format designed for rapid batch submission (and retrieval) of data. A single SOFT file can hold both data tables and accompanying descriptive information for multiple platforms, samples, and/or series records. SOFT files may be produced readily from common database and spreadsheet applications and can be uploaded directly to the database.

FTP deposit: If data are already in matrix format (e.g., Affymetrix pivot file), submission via a SOFT-formatted spreadsheet is recommended. Valid MAGE-ML-formatted (Spellman *et al.*, 2002) reports are also acceptable. These file types are transferred to GEO via FTP.

Full instructions and examples of these various submission routes and formats are provided on the GEO web site. All submissions are reviewed and checked by a GEO curator, ensuring that records contain meaningful information and are organized correctly. If no structural or content problems are identified the submissions are approved and assigned GEO accession numbers. If problems are identified, the curator will work with the submitter to make any modifications necessary to achieve successful deposit. The GEO accession numbers are unique and stable and may be quoted in corresponding manuscripts. The records may remain private for several months, typically pending manuscript publication. Submitters may generate a secondary account that enables collaborators or reviewers read-only, confidential access to prepublication data. Submitters retain full editorial control over their records and may perform updates and edits at any time.

Navigating GEO and Finding What You Need

Browsing

Original submitter-supplied platform, sample, and series records may be browsed using the repository browser at http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi. These browser pages allow data to be sorted by various categories, such as submitter, organism, platform and sample type, titles, release dates, and supplementary file type. DataSet records may be browsed at http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi and may be sorted by title, organism, type, creation date, and platform. Within records, reciprocal links are provided to all related records for easy, uninterrupted browsing.

Downloading

Several download options are available.

- Each platform, sample and series record has a mechanism at the head of the page that enables download (SOFT format) or viewing (HTML) of that record and/or related records, with the option to restrict to only descriptive data or tabular data.
- DataSet records include a link for download of a text tab-delimited value matrix and associated platform element gene annotation.
- All platform, sample, series, DataSet, and supplementary data are available for bulk download via FTP at ftp://ftp.ncbi.nih.gov/pub/geo/.

Query and Analysis

GEO provides a variety of strategies for locating and visualizing information of interest. Query approaches include standard and Boolean text-based searches, sequence-based searches, mining based on expression behavior characteristics, or combinations of these parameters. Figure 4 depicts a schematic overview of the query workflow and how the various features and tools are interlinked. A summary of where these features are located, their purpose, and methodology is provided.

Deciding where to begin a search generally depends on what type of information one needs to retrieve. Often, there is more than one way to identify relevant data. Users should always keep in mind that the features provided on the GEO site are not intended for robust systematic analyses. The heterogeneous nature of GEO data, coupled with the limitations of web browsing, limits to some extent the statistical tools that can be developed. Diverse data are treated similarly; criteria such as sample size, number of

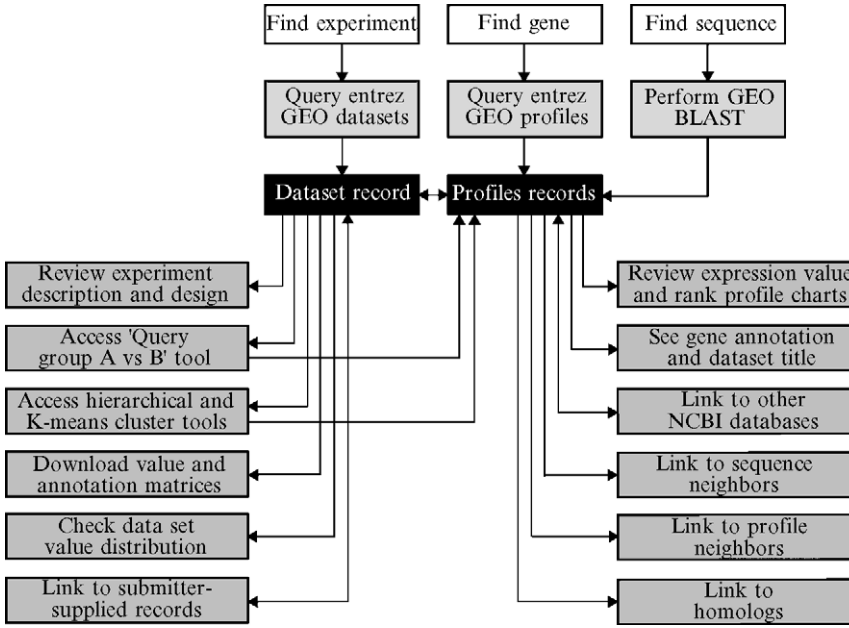

FIG. 4. Schematic overview of the query workflow and how the various features and tools are interlinked.

repeats, prior filtering, and normalization factors are not considered. That said, these tools are extremely useful for the quick and easy identification of relevant and noteworthy data.

Entrez GEO DataSets

Where: From the GEO home page or at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds.

Purpose: A query interface that facilitates identification of DataSets relevant to a particular area of study.

Method: Effective query and mining is achieved using keywords or Boolean phrases restricted to supported attribute fields (Table I). Retrievals display the DataSet titles, a brief experiment description, and a link to the complete DataSet record (Fig. 2), as well as links to publications and other databases.

Entrez GEO Profiles

Where: From the GEO home page or at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo.

TABLE I
ENTREZ QUALIFIER FIELDS[a]

Field name	Field description
GEO DataSets	
Author	Authors associated with the experiment
Experiment type	Experiment type, e.g., cDNA, genomic, protein, SAGE
GDS text	DataSet description text
GEO accession	GEO accession number
GEO description/title text	Text provided in the description/title of original records
Number of samples	Number of samples in the DataSet[b]
Number of platform probes	Number of platform reporters in the DataSet[b]
Organism	Organism from which the reporters on the array were derived/designed
Reporter identifier	Identifier for the array reporter (GenBank accession, gene name, etc.).
Sample source	Source biological material of the sample
Sample title	Sample title
Submitter institute	Submitter institute
Subset description	Description of the experimental variable
Subset variable type	Type of experimental variable, e.g., age, strain, gender
GEO profiles	
Experiment type	Experiment type, e.g., cDNA, genomic, protein, SAGE
Flag information	Specific experimental variable flags, e.g., age, strain, gender
Flag type	Flag types, e.g., rank and value subset effects
GDS text	DataSet description text
GEO accession	GEO accession number
GEO description/title text	Text provided in the description/title of original records
GI	Mapped GenBank identifier
Gene description	Gene description, symbol, alias
ID_REF	Unique identifier for a reporter as given on the array
Max value rank	Maximum value rank[b]
Max value in profile	Maximum value in profile[b]
Median value in GDS	Median value in DataSet[b]
Median value in profile	Median value in profile[b]
Min value rank	Minimum value rank[b]
Min value in profile	Minimum value in profile[b]
Number of samples	Number of samples in the DataSet[b]
Organism	Organism from which the samples were derived
Ranked standard deviation	Ranked standard deviation
Reporter identifier	Identifier for a reporter
Sample source	Source biological material of the sample

[a] Useful qualifier fields for performing restricted GEO DataSets and GEO profile queries.
[b] Possible range operation, e.g., 20:50[number of samples] will find DataSets containing 20 to 50 samples.

Purpose: A query interface that facilitates identification of gene expression profiles of interest.

Method: Effective query and mining is achieved using keywords or Boolean phrases restricted to supported attribute fields (Table I). Retrievals display the mapped gene name, the DataSet title, a thumbnail image of the gene expression profile, as well as links to publications and other databases. Clicking on the thumbnail image will enlarge the chart to display the full profile details and Sample subset partitions that reflect experimental design (Fig. 3).

Advanced Entrez Features

Where: The tool bar at the head of all NCBI Entrez query and retrieval pages.

Purpose: Facilitates powerful mining and linking across many NBCI databases (Schuler *et al*., 1996; Wheeler *et al*., 2005).

Method: The ''Preview/Limits'' link assists greatly in the construction of complex queries. Users employ indices to browse and/or select the terms by which data are described and build multipart queries. The ''History'' tab stores previous queries, which can be combined to form a new search query, enabling sophisticated mining that traverses DataSets and platforms. The ''Display'' pull-down menu enables users to find related data in other Entrez resources in batch mode.

DataSet Clusters

Where: On the DataSet record under the ''analysis'' button.

Purpose: Clustering is a popular method used to visualize and examine high-dimensional DataSets. Typically, the goal of a microarray cluster analysis is to organize genes so that those with similar expression patterns are grouped together. It can be hypothesized that genes that behave similarly might have a coordinated transcriptional response, possibly inferring a common function or regulatory elements.

Method: Many different clustering algorithms exist (see Gollub and Sherlock, 2006); all employ various combinations of mathematical distance metrics and linkages (Eisen *et al*., 1998). Nine varieties of precomputed hierarchical clusters are available on GEO DataSet records, as well as user-defined K-means or K-median clustering. Results are depicted as a color-coded ''heat map'' image, where rows represent individual elements on the array (genes) and columns represent individual samples (hybridizations), and color A (high expression level) transitions into color B (low expression level). Users

can scan these images visually for cluster ''hot spots'' that represent a group of genes with similar expression. The heat maps are interactive; after selecting a region, or regions, of interest using a movable box, corresponding data may be downloaded as a text file or linked to genes in Entrez GEO profiles. Care must be taken not to over-interpret cluster output. Different clustering algorithms may yield different clustering solutions using the same data. Clustering provides suggestions for possible relationships between data, but does not prove them.

Profile Neighbors

Where: The ''Profile Neighbors'' link on the top right side of Entrez GEO profile retrievals.

Purpose: Connects groups of genes that show a similar or reversed profile shape within a DataSet. It can be hypothesized that genes that behave similarly might be coregulated or have related functionality.

Method: Profile neighbors are precalculated using an adjusted Pearson linear correlation. The user need only click the ''Profile Neighbors'' link to retrieve related genes. Currently, Profile neighbors are subject to a GEO-defined arbitrary cutoff limit imposed in order to restrict the number of links that can be managed effectively.

Sequence Neighbors

Where: The ''Sequence Neighbors'' link on the top right side of Entrez GEO profile retrievals.

Purpose: Connects groups of genes related by nucleotide sequence similarity across all DataSets. Genes related by sequence similarity can provide insights into the possible function of the original sequence if it has not yet been characterized or can identify related gene family members.

Method: Sequence neighbors are precalculated using standard BLAST (Altschul *et al.*, 1990). The user need only click the ''Sequence Neighbors'' link to retrieve related genes. Currently, Sequence neighbors are subject to a GEO-defined arbitrary cutoff limit imposed in order to restrict the number of links that can be managed effectively.

Links

Where: The ''Links'' link on the top right side of Entrez GEO profiles and Entrez GEO DataSets retrievals.

Purpose: Connects GEO data to related data in other NCBI resources, facilitating seamless navigation and cross-referencing between multiple data domains.

Method: Where possible, reciprocal links are provided to and from GenBank, PubMed, Gene, UniGene, OMIM, Homologene, Taxonomy, SAGEMap, and MapViewer databases. The user need only click the ''Links'' link and select the relevant resource from the pull-down menu to link to retrieve related data.

Geo Blast

Where: The GEO BLAST link on the GEO home page.

Purpose: Retrieves gene profiles that are related to a user-defined nucleotide sequence of interest.

Method: This tool performs a BLAST (Altschul *et al.*, 1990) search of a user-provided nucleotide sequence against all GenBank identifiers represented on microarray platforms or SAGE libraries in GEO. Retrievals resemble conventional BLAST output with each alignment receiving a score and expected value and a link to corresponding GEO profiles. This interface is helpful in locating expression data for specified nucleotide sequences, for identifying sequence homologs, for example, related gene family members or for cross-species comparisons, or for providing insight into potential roles of the original sequence if it has not yet been characterized functionally.

Sorting and Limit Options Using Subset Effects Flags

Where: Intrinsic to standard Entrez GEO profiles retrievals, which are default ordered according to subset effect flags and specifiable using [Flag Type] and [Flag information] qualifiers (Table I) in Entrez GEO profiles.

Purpose: Attempts to identify genes that display marked differences in expression level according to experimental variables.

Method: Genes whose values or ranks pass a threshold of statistical difference between any nonsingle experimental variable subset and another are flagged in the database. This allows users to search across all GEO for genes that show an interesting effect with respect to particular experimental variable types, such as ''age.'' The fact that standard Entrez GEO profile retrievals are default ordered according to these flags makes potentially interesting results more visible (alternative sorting options include profile deviation and mean value). It is important to realize that subset effects are calculated with arbitrarily defined thresholds with no consideration of data type and processing and merely provide suggestions of what could be interesting profiles.

Query Group A vs B Tool

Where: On the DataSet record on the right side of the subset assignment section.

Purpose: Assists filtering and identification of gene profiles that display marked differences in the expression level between two specified sets of samples within a DataSet.

Method: Using checkboxes, the user assigns one or more samples to group A and other samples to group B. Samples are selected/deselected on the basis of their experimental subset designations. The user then chooses from several varieties of filtering procedures and stringency parameters by which to compare the two groups, including one-tailed or two-tailed *t* tests or a mean log values or ranks fold difference. Genes that meet the user-defined criteria are presented in Entrez GEO profiles. Note that this tool uses rudimentary means of filtering data, as retrievals may have no statistical significance; the compared subsets may be too small to provide any statistic value.

Conclusion

DNA microarray technology has led to a rapid accumulation of gene expression data. GEO serves as a unifying resource for these data, operating primarily as a public archive, but also providing flexible data mining strategies and tools that allow users to query, filter, select, and inspect data in the context of their specific interests. Many of these features use traditional data reduction techniques designed to filter inherently noisy data and concise displays that allow human scanning. The integration of GEO data with extensive sequence, mapping, and bibliographic resources via the Entrez system of linked databases offers additional ancillary information that can assist in the interpretation of biological data and evaluate the relevance of microarray results.

Examination of published gene expression data can help researchers prioritize candidates for further study and direct the design of new experiments. The literature reveals that researchers are using GEO data to complement and support their own studies (e.g., Brockington *et al.*, 2005; Nakai *et al.*, 2005; Ozyildirim *et al.*, 2005; Rico-Bautista *et al.*, 2005; Yant *et al.*, 2005).

Compiling large volumes of diverse gene expression data into one location and making them accessible through common integrated interfaces impart a powerful investigative factor not attainable when considering solitary experiments. This large compendium of data affords more opportunity to gather corroboratory evidence for global metabolic and

regulatory networks, to investigate what the majority of evidence implies about the behavior and function of a gene or group of genes, and to generate hypotheses on functional models and themes (e.g., Jordan *et al.*, 2004; Ott *et al.*, 2005; Zhou *et al.*, 2005). This macro approach to discovery will only strengthen as the database continues to grow.

Because the GEO database and tools continue to undergo intensive development and modification, the features and data presentation strategies discussed in this chapter will evolve over time. To receive announcements of site developments, subscribe to the GEO-announce list at geo@ncbi.nlm.nih.gov.

Acknowledgments

The authors unreservedly acknowledge the efforts of the GEO curation and programming staff, including Tugba Suzek, Dennis Troup, Steve Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, and Alexandra Soboleva. Also, Todd Groesbeck is thanked for assistance with manuscript figures. This chapter is an official contribution of the National Institutes of Health; not subject to copyright in the United States.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S. A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J., and Winegarden, N. (2004). Submission of microarray data to public repositories. *PLoS Biol.* **2**(9), e317.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: Mining millions of expression profiles—database and tools. *Nucleic Acids Res.* **33,** D562–D566.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genet.* **29,** 365–371.

Brazma, A., Kapushesky, M., Parkinson, H., Sarkins, U., and Shojatalab, M. (2006). Data storage and analysis in ArrayExpress. *Methods Enzymol.* **411,** 370–386.

Brockington, M., Torelli, S., Prandini, P., Boito, C., Dolatshad, N. F., Longman, C., Brown, S. C., and Muntoni, F. (2005). Localization and functional analysis of the LARGE family of glycosyltransferases: Significance for muscular dystrophy. *Hum. Mol. Genet.* **14**(5), 657–665.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.

Gollub, J., and Sherlock, G. (2006). Clustering microarray data. *Methods Enzymol.* **411** (this volume).

Gonzalez, R., Yang, Y. H., Griffin, C., Allen, L., Tigue, Z., and Dobbs, L. (2005). Freshly isolated rat alveolar type I cells, type II cells, and cultured type II cells have distinct molecular phenotypes. *Am. J. Physiol. Lung Cell Mol. Physiol.* **288**(1), L179–L189.

Jordan, I. K., Marino-Ramirez, L., Wolf, Y. I., and Koonin, E. V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**(11), 2058–2070.

Nakai, H., Wu, X., Fuess, S., Storm, T. A., Munroe, D., Montini, E., Burgess, S. M., Grompe, M., and Kay, M. A. (2005). Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. *J. Virol.* **79**(6), 3606–3614.

Ott, S., Hansen, A., Kim, S. Y., and Miyano, S. (2005). Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution. *Bioinformatics* **21**(2), 227–238.

Ozyildirim, A. M., Wistow, G. J., Gao, J., Wang, J., Dickinson, D. P., Frierson, H. F., Jr., and Laurie, G. W. (2005). The lacrimal gland transcriptome is an unusually rich source of rare and poorly characterized gene transcripts. *Invest. Ophthalmol. Vis. Sci.* **46**(5), 1572–1580.

Rico-Bautista, E., Greenhalgh, C. J., Tollet-Egnell, P., Hilton, D. J., Alexander, W. S., Norstedt, G., and Flores-Morales, A. (2005). Suppressor of cytokine signaling-2 deficiency induces molecular and metabolic changes that partially overlap with growth hormone-dependent effects. *Mol. Endocrinol.* **19**(3), 781–793.

Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* **266,** 141–162.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr., and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3,** RESEARCH0046.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pontius, J. U., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **33,** D39–D45.

Yant, S. R., Wu, X., Huang, Y., Garrison, B., Burgess, S. M., and Kay, M. A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell. Biol.* **25**(6), 2085–2094.

Zhou, X. J., Kao, M. C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E., and Wong, W. H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature Biotechnol.* **23**(2), 238–243.